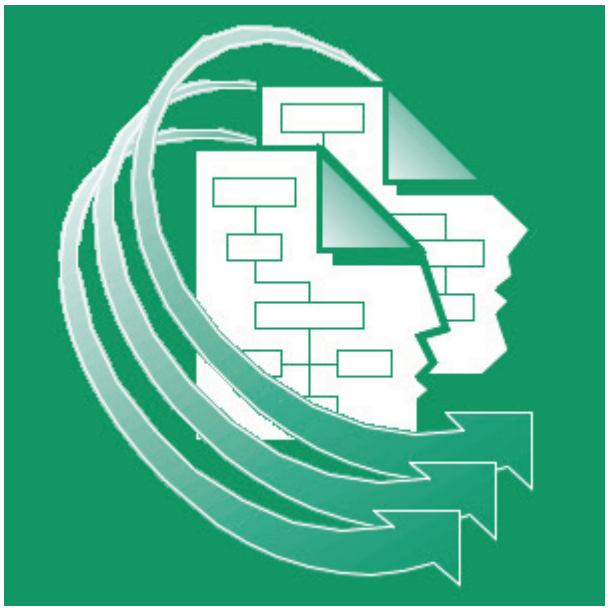
Business and Labour Market Analysis Division & Labour Statistics Division

Guide to the Analysis of

Workplace and Employee Survey 2000



[Version 2.1: September 2002]

24 R.H. Coats Building
Ottawa, Ontario
K1A 0T6
(CANADA)

Tel: (613) 951-4233
Fax: (613) 951-4087
e-mail: fangtao@a.statcan.ca

www.statcan.ca

Statistics Canada Statistique Canada

Canada

Note to Research Data Centre Users: Several identification variables will not be available in the Research Data Centres in order to protect respondent confidentiality. Variables which are not available in the Research Data centres will be shown with an (HO) for Head Office in the data dictionary.

1 – QUICK START GUIDE

This quick start guide is intended to give experienced microdata users the information they need to begin accessing Workplace and Employee Survey data. The following links provide the necessary information to get started. Please read the notes that follow the links to ensure proper use and interpretation of the data.

Electronic Data Dictionary Questionnaires

1. **Use the survey weights in all analyses.** The employer survey is based on a stratified sample design that incorporates information on region, industry and employment size. Employees are selected randomly within each sampled business location. The sample is not "self representing" and failure to use the weights will result in estimates that do not relate to a known population. To those familiar with the term, we are strong advocates of "design based estimation".

2. **Use the appropriate survey weights.** There are three sets of survey weights available for both the 1999 and 2000 data: employer weights, employee weights and employer-linked weights. The reasons for the first two sets of weights are obvious, studies can be carried out independently at both the employer and employee level of the WES. However there were a number of locations from which we received employer responses, but no employee responses. These 'voids' are, of course, built into the employee weights, but necessitate a separate set of weights (the employer-linked weights) for employer-side analyses that incorporate employee characteristics. The 2000 weights should be used in all longitudinal analyses.

   Research data centre users refer to Appendix 7 section on bootstrap weighting.

3. **Account for the survey design in variance calculations.** Even though the use of the appropriate survey weights will result in consistent estimates, most software packages will underestimate the variance of the estimates because they do not account for the design of the survey. In Appendix 7, we describe how to calculate correct variances (or reasonable approximations) in several different ways. Calculating an appropriate variance is the only way to determine the precision of the estimates and relationships that support your analyses.

4. **Choose an appropriate model for linked analyses**. Combining variables from both the employer and employee surveys will enhance many analyses and open new avenues of research, however such linked studies will require careful selection of the statistical model. Multi-level data will not conform to the assumptions of most simple statistical models. Some of the appropriate techniques are briefly discussed in Appendix 5. A bibliography of more detailed applications of these techniques is also included.

5 – Micro data Files with a prefix 'Im' for data and 'ei' for edit and imputation levels.

6 – Dummy Data Files with a prefix 'Dm' for data and 'de' for edit and imputation levels.

7 – Macro estimates - Control Totals Files with a prefix 'Ma'.

All the files mentionned in 5-7 are available in SAS, SPSS and STATA.

APPENDIX 1

Introduction

Why have a Linked Workplace and Employee Survey?

Advanced economies are constantly evolving. The key stimuli for this evolution are new technologies (particularly information technologies), increasing international competition and the continued expansion of transnational enterprises. Firms respond in a number of ways: increasingly embracing new technologies; re-organizing or re-engineering their workforces; or resorting to downsizing or other elements of numerical flexibility. For firms, these trends create challenges in the management and development of human resources. For policy-makers, education and training are central policy prescriptions for increasing prosperity.

In this evolving environment, firms are thought to have undergone dramatic change in the areas of technology adoption, organizational change, training patterns, business strategies, levels of competition, and the manner in which they engage labour. Workers, on the other hand, experience this evolution through changes in job creation rates, job stability, wages and wage inequality, training, the use of advanced technologies, and the type of employment contracts available.

Due to a well-developed set of household (worker) surveys, we in Canada have a good understanding of workers' outcomes regarding wages and wage inequality, job stability and layoffs, training, job creation, and unemployment. What has been missing on the employees' side is the ability to link these changes to events taking place in firms. Such a connection is necessary if we hope to understand the association between labour market changes and demand-side pressures, which stem from global competition, technological change, and the drive to improve human capital, among other things. Thus, one primary goal of the WES is to establish a link between events occurring in establishments and the outcomes for workers.

The advantage of a linked survey is depicted in Figure 1. This chart displays the main content blocks in the two surveys. Note that there is reference to establishment and worker outcomes. Analysis of these events can be informed not only by the characteristics of the establishment -- as has been done in other firm surveys -- but also by the characteristics of the workers. Similarly, worker outcomes can be informed not only by data on the workers themselves, as has always been the case, but also by new establishment data.

For example, this link allows changes in the levels and distributions of wages of workers to be associated with events occurring in establishments, such as the adoption of technology, or competing in international markets. Much of the earnings inequality literature suggest that technology and rising international trade are major contributors to inequality. Research on many other labour market issues would be enhanced by the existence of such a link. Issues that have formerly been considered primarily from the supply side, often within the context of a human capital model, could be viewed increasingly from the demand side of the labour market. This might include issues such as job stability, the determinants of wages, the creation and destruction of different types of jobs, training levels among different types of workers, etc.

The establishment-worker link also contributes to improved measurement of a number of establishment–level variables. The characteristics of an establishment's workforce are often an important determinant of the behavior of a firm. However, data on workforce characteristics have been lacking or poorly measured in establishment surveys. The WES allows establishment variables -- such as training incidence and intensity, occupational and educational distribution of the workforce, use of technology by the workers, various workplace practices such as quality circles, fringe benefit levels, the distribution of wages,

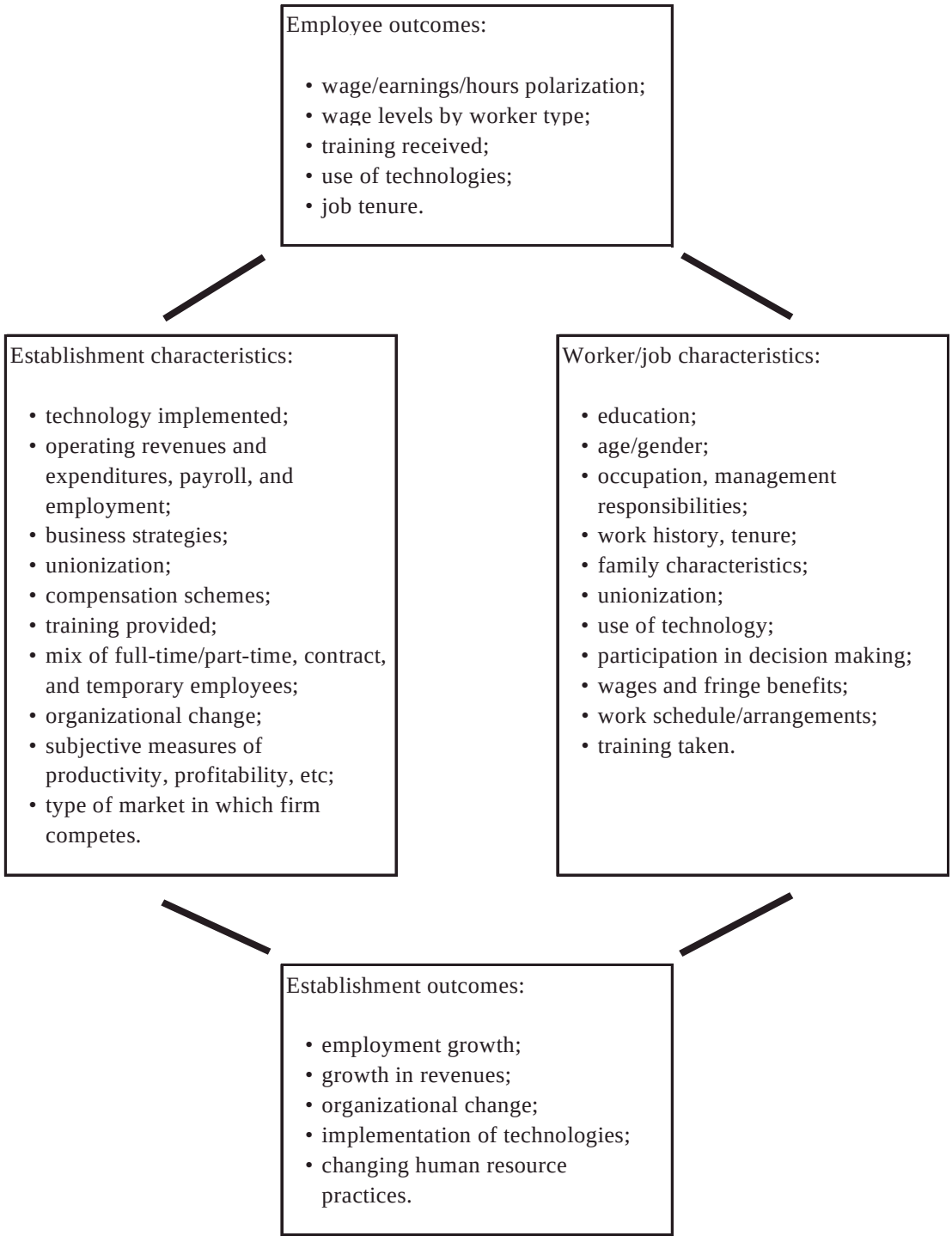
and a host of others -- to be better measured than in the past. Workers can provide more reliable and detailed data on these variables than can establishment level respondents.

The second goal of the survey is to develop a better understanding of what is indeed occurring in companies in an era of substantial evolution. Just how many companies have implemented new information technologies? On what scale? What kind of training is associated with this? What type of organizational change is occurring in firms? What types of business strategies are firms relying on to thrive during this period of change, and do they vary dramatically across firms? How important are human resource development activities and strategies, or are they largely ignored by most establishments? Do firms that adopt one set of strategies in fact adopt many (e.g., adoption of technologies, innovation, human resource development, and organizational changes)? Is there a set of high-performance workplaces that tend to move on many fronts? These are the kinds of issues addressed in the WES.

While the available household surveys inform us about significant labour market changes, there has not been a corresponding set of establishment surveys that deal with new concerns. Some limited survey work has been done. The WES is an attempt to extend this in the context of a general worker–workplace survey.

Finally, the third objective is to extend surveying infrastructure. To a considerable extent WES is seen as the development of the infrastructure necessary to conduct integrated establishment-household surveys. Core content will be repeated over successive waves of the survey, while content covering less frequent events will be cycled out in alternative waves. Once response burden and data quality has been assessed across several waves, new content could be cycled in to meet changing information needs.

Figure 1: The Link Between the Workplace Survey Content, Employee Survey Content, and Outcomes



APPENDIX 2

Concepts and Definitions

OBJECTIVES

The Workplace and Employee Survey (WES) is designed to explore a broad range of issues relating to employers and their employees. The survey aims to shed light on the relationships among competitiveness, innovation, technology use and human resource management on the employer side and technology use, training, job stability and earnings on the employee side.

The survey is unique in that employers and employees are linked at the micro data level; employees are selected from within sampled workplaces. Thus, information from both the supply and demand sides of the labour market is available to enrich studies on either side of the market.

Sample sizes and Response rates

WES was conducted for the first time during the summer (employer survey part) and fall of 1999 (employee survey part). The employer sample is longitudinal – the sampled locations will be followed over time, with the periodic addition of samples of new locations to maintain a representative cross section. Employees will be followed for two years only, due to the difficulty of integrating new employers into the location sample as workers change companies. As such, fresh samples of employees will be drawn on every second survey occasion (i.e. first, third, fifth). This longitudinal aspect will allow researchers to study both employer and employee outcomes over time in the evolving workplace.

A2.1 **Sample Sizes and Estimated Populations 1999**

Industry / Workplace size / Region	Workplaces		Employee	
	Number of respondents	**Estimated population**	**Number of respondents**	**Estimated population**
Overall	**6,322**	**718,083**	**20,167**	**10,626,280**
Industry				
Forestry, mining, oil and gas extraction	292	12,610	1,100	185,179
Labour intensive tertiary manufacturing	408	20,170	1,556	496,863
Primary product manufacturing	320	7,263	1,392	398,708
Secondary product manufacturing	293	11,932	1,143	367,268
Capital intensive tertiary manufacturing	359	16,191	1,429	584,255
Construction	608	57,736	2,021	420,546
Transportation, warehousing and wholesale trade	711	88,715	2,782	1,109,092
Communication and other utilities	421	9,740	1,326	243,785
Retail trade and consumer services	524	234,636	1,764	2,593,009
Finance and insurance	506	36,543	1,841	505,794
Real estate, rental and leasing operations	364	27,610	1,098	182,695
Business services	468	79,010	1,728	1,000,274
Education and health services	704	100,198	2,986	2,339,685
Information and cultural industries	344	15,729	1,374	350,391
Workplace size				
1-19 employees	2,789	626,933	5,607	3,408,392
20-99 employees	1,711	77,560	7,780	2,971,669
100-499 employees	1,300	11,781	6,672	2,167,271
500 employees or more	522	1,810	3,481	2,230,211
Region				
Atlantic	774	62,542	2,892	711,924
Quebec	1,427	155,335	5,510	2,570,035
Ontario	1,577	260,983	5,781	4,295,566
Manitoba	420	25,651	1,556	409,578
Saskatchewan	342	28,782	1,221	328,707
Alberta	852	81,062	3,089	1,107,662
British Columbia	930	103,729	3,491	1,354,071
			1,556	409,578

A2.2 **Response Rates 1999**

	Workplace response rate (%)	**Employee response rate (%)**
Overall	**94.0**	**83.1**

A2.3 **Sample Sizes and Estimated Populations 2000**

Industry / Workplace size / Region	Workplaces		Employee	
	Number of respondents	**Estimated population**	**Number of respondents**	**Estimated population**
Overall	**6,068**	**668,188**	**20,167**	**10,626,280**
Industry				
Forestry, mining, oil and gas extraction	278	11,580	970	192,089
Labour intensive tertiary manufacturing	389	18,906	1,299	497,873
Primary product manufacturing	306	6,959	1,221	398,154
Secondary product manufacturing	275	11,631	961	368,917
Capital intensive tertiary manufacturing	344	15,521	1,225	557,711
Construction	576	49,848	1,681	402,466
Transportation, warehousing and wholesale trade	687	81,136	2,367	1,111,175
Communication and other utilities	394	9,053	1,142	245,309
Retail trade and consumer services	540	220,991	1,538	2,585,846
Finance and insurance	485	34,613	1,621	511,809
Real estate, rental and leasing operations	325	22,945	842	175,715
Business services	460	76,742	1,462	1,018,702
Education and health services	680	93,833	2,652	2,339,542
Information and cultural industries	329	14,428	1,186	349,721
Workplace size				
1-19 employees	2,604	574,241	4,901	3,398,935
20-99 employees	1,687	80,388	6,619	3,071,802
100-499 employees	1,280	11,763	5,724	2,171,798
500 employees or more	497	1,797	2,923	2,112,495
Atlantic	746	59,071	2,578	740,971
Quebec	1,365	141,823	4,525	2,422,468
Ontario	1,529	251,441	4,983	4,356,308
Manitoba	400	21,829	1,375	408,677
Saskatchewan	323	26,025	1,091	338,520
Alberta	821	76,225	2,602	1,111,862
British Columbia	884	91,774	3,013	1,376,223

A2.4 **Response Rates 2000**

	Workplace response rate (%)	**Employee response rate (%)**
Overall	**94.0**	**87**

Target population

The target population for the employer component is defined as all business locations operating in Canada that have paid employees in March , with the following exceptions:

Employers in Yukon, Nunavut and Northwest Territories

Employers operating in crop production and animal production; fishing, hunting and trapping; private households, religious organizations and public administration.

The target population for the employee component is all employees working or on paid leave in March in the selected workplaces who receive a Customs Canada and Revenue Agency T-4 Supplementary form. If a person receives a T-4 slip from two different workplaces, then the person will be counted as two employees on the WES frame.

Survey Population

The survey population is the collection of all units for which the survey can realistically provide information. The survey population may differ from the target population due to operational difficulties in identifying all the units that belong to the target population.

WES draws its sample from the Business Register (BR) maintained by the Business Register Division of Statistics Canada, and from lists of employees provided by the surveyed employers.

The Business Register is a list of all businesses in Canada, and is updated each month using data from various surveys, profiling of businesses and administrative sources.

Applicable Population

Workplace
The applicable population follows the flow of the questionnaire and represents the estimated population of workplaces based on our sample.

Employee
The applicable population follows the flow of the questionnaire and represents the estimate population of employees based on our sample.

Reference Period

There are two reference periods used for the WES. Questions concerning employment breakdown use the last pay period of March for the reference year while other questions refer to the last 12-month period ending in March of the reference year.

Sample Design

The survey frame is a list of all locations that carries contact and classification (e.g., industrial classification) information on the units. This list is used for sample design and selection; ultimately, it provides contact and classification information for the selected units.

Workplace Survey

The survey frame for the Workplace component of WES was created from the information available on the Statistics Canada Business Register.

Prior to sample selection, the business locations on the frame were stratified into relatively homogeneous groups called *strata*, which were then used for sample allocation and selection. The WES frame was stratified by industry (14), region (6), and size (3), which was defined using estimated employment. The size stratum boundaries were typically different for each industry/region combination. The cut-off points defining a particular size stratum were computed using a model-based approach. The sample was selected using Neyman allocation. This process generated 252 strata with 9,144 sampled business locations.

All sampled units were assigned a sampling weight (a raising factor attached to each sampled unit to obtain estimates for the population from a sample). For example, if two units were selected at random and with equal probability out of a population of ten units, then each selected unit would represent five units in the population, and it would have a sampling weight of five.

As an example, the inaugural WES survey collected data from 6,322 out of the 9,144 sampled employers. The remaining employers were a combination of workplaces determined to be either out-of-business, seasonally inactive, holding companies, or out-of-scope. The majority of non-respondents were owner-operators with no paid help and in possession of a payroll deduction account.

Employee Survey

The frame for the employee component of WES was based on lists of employees made available to interviewers by the selected workplaces. A maximum of twelve employees was sampled using a probability mechanism. In workplaces with fewer than four employees, all employees were selected.

Data Collection

Data collection, data capture, preliminary editing and follow-up of non-respondents were all done in Statistics Canada Regional Offices. In 1999, workplace data were collected through personal interviews. In 2000, computer assisted telephone interviews were conducted. For about 20% of the surveyed units (mostly large workplaces), more than one contact person was required. For the employee component, telephone interviews were conducted with persons who had agreed to participate in the survey by filling out and mailing in an employee participation form.

Statistical Edit and Imputation

Following collection, all data were analyzed extensively. Extreme values were listed for manual inspection in order of priority determined by the size of the deviation from average behaviour and the size of their contribution to the overall estimate.

Respondents who opted not to participate in the survey – *total non-response* – were removed and the weights of the remaining units were adjusted upward to preserve the representativity of the sample. For respondents who did not provide all required fields – *item non-response* – a statistical technique called *imputation* was used to fill in the missing values for both employers and employees.

The WES components were treated independently even if some questions on the employee questionnaire could have been imputed from the related workplace questionnaire.

Estimation

The reported (or imputed) values for each workplace and employee in the sample are multiplied by the weight for that workplace or employee; these weighted values are summed up to produce estimates. An initial weight equal to the inverse of the original probability of selection is assigned to each unit. To calculate variance estimates, the initial survey weights are adjusted to force the estimated totals in each industry/region group to agree with the known population totals. These adjusted weights are then used in forming estimates of means or totals of variables collected by the survey.

Variables for which population totals are known are called auxiliary variables. They are used to calibrate survey estimates to increase their precision. Each business location is calibrated to known population totals at the industry/region level. The auxiliary variable used for WES is total employment obtained from the Survey of Employment, Payrolls and Hours.

Estimates are computed for many domains of interest such as industry and region.

Data Quality

Coefficient of variation rules

Estimates with a coefficient of variation greater than 33.5 percent are not published.

Estimates with a coefficient of variation in the range of 25 to 33.5 percent are published with a cautionary flag, denoting their relatively high variability.

Any survey is subject to errors. While considerable effort is made to ensure a high standard throughout all survey operations, the resulting estimates are inevitably subject to a certain degree of error. Errors can arise due to the use of a sample instead of a complete census, from mistakes made by respondents or interviewers during the collection of data, from errors made in keying in the data, from imputation of a consistent but not necessarily correct value, or from other sources.

Sampling Errors

The true sampling error is unknown; however, it can be estimated from the sample itself by using a statistical measure called the *standard error*. When the standard error is expressed as a percent of the estimate, it is known as the relative standard error or *coefficient of variation*.

Non-Sampling Errors

Some non-sampling errors will cancel out over many observations, but systematically occurring errors (i.e. those that do not tend to cancel) will contribute to a bias in the estimates. For example, if respondents consistently tend to underestimate their sales, then the resulting estimate of the total sales will be below the true population total. Such a bias is not reflected in the estimates of standard error. As the sample size increases, the sampling error decreases. However, this is not necessarily true for the non-sampling error.

Coverage Errors

Coverage errors arise when the survey frame does not adequately cover the target population. As a result, certain units belonging to the target population are either excluded (under-coverage), or counted more than once (over-coverage). In addition, out-of-scope units may be present on the survey frame (over-coverage).

Response Errors

Response errors occur when a respondent provides incorrect information due to misinterpretation of the survey questions or lack of correct information, gives wrong information by mistake, or is reluctant to disclose the correct information. Gross response errors are likely to be caught during editing, but others may simply go through undetected.

Non-response Errors

Non-response errors can occur when a respondent does not respond at all (total non-response) or responds only to some questions (partial non-response). These errors can have a serious impact on estimates if the non-respondents are systematically different from the respondents in survey characteristics and/or the non-response rate is high.

Processing Errors

Errors that occur during the processing of data represent another component of the non-sampling error. Processing errors can arise during data capture, coding, editing, imputation, outlier treatment and other types of data handling. A coding error occurs when a field is coded erroneously because of misinterpretation of coding procedures or bad judgement. A data capture error occurs when data are misinterpreted or keyed in incorrectly.

Joint Interpretation of Measures of Error

The measure of non-response error and the coefficient of variation must be considered jointly to assess the quality of the estimates. The lower the coefficient of variation and the higher the response fraction, the better will be the published estimate.

Confidentiality

The information presented in this publication has been reviewed to ensure that the confidentiality of individual responses is respected. Any estimate that could reveal the identity of a specific respondent is declared confidential, and consequently not published.

Response/Non-response

a) ***Response rate:*** includes all units, which responded by providing "usable information" during the collection phase.

b) ***Refusal rate:*** includes those units, which were contacted but refused to participate in the survey.

Industry Definitions

WES industry codes	Industry descriptions	3-digit North American Industry Classification System (NAICS)
01	Forestry / mining / oil and gas extraction	113, 115, 211, 212, 213
02	Labour intensive tertiary manufacturing	311, 312, 313, 314, 315, 316, 337, 339
03	Primary product manufacturing	321, 322, 324, 327, 331
04	Secondary product manufacturing	325, 326, 332
05	Capital intensive tertiary manufacturing	323, 333, 334, 335, 336
06	Construction	231, 232
07	Transportation / warehousing / wholesale trade	411, 412, 413, 414, 415, 416, 417, 418, 419, 481, 482, 483, 484, 485, 486, 487, 488, 493
08	Communication and other utilities	221, 491, 492, 562
09	Retail trade & consumer services	441, 442, 443, 444, 445, 446, 447, 448, 451, 452, 453, 454, 713, 721, 722, 811, 812
10	Finance and insurance	521, 522, 523, 524, 526
11	Real estate, rental, leasing operations	531, 532
12	Business services	533, 541, 551, 561
13	Education and health services	611, 621, 622, 623, 624
14	Information and cultural industries	511, 512, 513, 514, 711, 712

Industrial activities excluded from WES	3-digit North American Industry Classification System (NAICS)
Crop production / animal production	111, 112
Fishing, hunting and trapping	114
Religious organizations	813
Private households	814
Federal government public administration	911
Provincial and territorial public administration	912
Local, municipal and regional public administration	913
Aboriginal public administration	914
International and other extra-territorial public administration	919

Occupation Definitions

A. Employee:

Any person receiving pay for services rendered in Canada or for paid absence, and for whom you are required to complete a Canada Customs and Revenue Agency T-4 Form.

Employee:

A. Full-time employee: An employee working 30 or more hours per week.
B. Part-time employee: An employee working less than 30 hours per week.
C. Permanent employee: An employee who has no set termination date.
D. Non-permanent employee: An employee who has a set termination date or an agreement covering the period of employment (e.g. temporary or seasonal).

B. Independent contractor:

A person providing products or services under contract with your location but for whom the completion of a Canada Customs and Revenue Agency T-4 Form is not required. This person may be an employee of another business or a home worker (e.g. computer consultant, piecework seamstresses, etc).

C. Management:

1. Managers

(a) Senior Managers

Include the most senior manager in the workplace and other senior managers whose responsibilities would normally span more than one internal department. Most small workplaces would only have one senior manager. Examples: president of single location company; retail store manager; plant manager; senior partners in business services firms; production superintendent; senior administrator in public services enterprise; as well as vice-presidents, assistant directors, junior partners and assistant administrators whose responsibilities cover more than one specific domain.

(b) Specialist Managers

Managers who generally report to senior management and are responsible for a single domain or department. This category would normally include assistant directors or the equivalent in small workplaces. Examples: department heads or managers (engineering, accounting, R&D, personnel, computing, marketing, sales, etc.); heads or managers of specific product lines; junior partners or assistant administrators with responsibilities for a specific domain; and assistant directors in small locations (without an internal department structure).

D. Non-Management:

2. Professionals

Employees whose duties would normally require at least an undergraduate university degree or the equivalent. Examples: medical doctors, lawyers, accountants, architects, engineers, economists, science professionals, psychologists, sociologists, registered nurses, marketing and market research professionals, nurse-practitioners and teaching professionals. Include computing professionals whose duties would normally require a minimum of an undergraduate degree in computer science. Include professional project managers and supervisors not included in senior managers (C.1 (a)) and specialist managers (C.1 (b)).

3. Technical / Trades

Composed of:

(a) Technical / Semi-professional workers

Employees whose duties would normally require a community college certificate /diploma or the equivalent and who are not primarily involved in the marketing /sales of a product or service. Examples: technologists, lab technicians, registered nursing assistants, audio-visual technicians; ECE-trained caregivers; technology trainers; physiotherapists; legal secretaries and draftspersons. Include computer programmers and operators whose duties would normally require a community college certificate or diploma. Include semi-professional project managers and supervisors not included managers (C.1) and professionals (D.1). Exclude marketing /sales personnel with non-university accreditation.

(b) Trades /Skilled production, operation and maintenance

Non-supervisory staff in positions requiring vocational /trades accreditation or the equivalent. Examples: construction trades, machinists, machine tenders, stationary engineers, mechanics, beauticians /barbers /hairdressers, butchers and repair occupations that do not normally require a post-secondary certificate or diploma.

4. Marketing / Sales

Non-supervisory staff primarily engaged in the marketing / sales of products or services. Examples: retail sales clerks, waiters/waitresses, telemarketers, real estate agents, insurance agents and loans officers. Exclude employees whose duties require a university degree and professional accreditation (professionals (D.1)), those whose duties require a community college certificate /diploma (technical/trades (D.2)) and those whose duties are primarily supervisory (managers (C.1)).

5. Clerical / Administrative

Non-supervisory staff providing clerical or administrative services for internal or external clients. Examples: secretaries, office equipment operators, filing clerks, account clerks, receptionists, desk clerks, mail and distribution clerks, bill collectors and claims adjusters. Duties do not normally require post-secondary education or responsibility for marketing or sales.

6. Production workers with no trade/certification, operation and maintenance

Non-supervisory staff in production or maintenance positions that require no vocational /trades accreditation or the equivalent in on-the-job training. Examples: assemblers, packers, sorters, pilers, machine operators, transportation equipment operators (drivers), warehousemen, and cleaning staff. As a rough guideline, jobs in this category require no more than a one-month training for someone with no trade or vocational accreditation.

7. Other

If you have a large number of employees who do not correspond to any of the above categories, please write in their occupation(s) in the space provided below.

Occupation Definitions

WES	SOC91
01 Managers	A011-A016; A111-A114; A121-A122; A131; A141; A211 A221-A222; A301-A303; A311-A312; A321-A324; A331-A334; A341-A343; A351-A353; A361; A371-A373; A381; A391-A392; E037
02 Professionals	B011-B014; B021-B022; B313; B315-B318; C011-C015; C021-C023; C031-C034; C041-C048; C051-C054; C061-C063; C111-C113; C121; C152; 162-C163; D011-D014; D021-D023; D031-D032; D041-D044; D111-D112; D211; D232; E011-E012; E021-E025; E031-E036; E038; E111-E112; E121; E130-E133; E211-E214; E216; F011-F013; F021-F025 F031-F034; F111; F121; F123; F143;
03 Technical/Trades	B111-B116; B212-B214; B311-B312; B314; B411-B415; B576; C122-C125; C131-C133; C141-C144; C151; C153-C155; C161; C164; C171-C175; D212-D219; D221-D223; D231; D233-D235; D311-D313; E215; F035-F036; F112; F122; F124-F127; F131-F132; F141-F142; F144-F145; F151-F154; G011-G016; G111; G121; G133-G134; G411-G412; G512; G611-G612; G621-G625; G631; G711-G712; G722; G812-G813; G911-G912; G921-G922; G933; G941-G942; G951; G981; H011-H019; H021-H022; H111-H113; H121-H122; H131-H134; H141-H145; H211-H217; H221-H222; H311-H312; H321-H325; H411-H418; H421-H422; H431-H435; H511-H514; H521-H523; H531-H535; H611-H612; H621-H623; H711-H714; H721-H722; H731; H736-H737; I011-I017; I021-I022; I111; I121-I122; I131-I132; I141-I142; I151; I161-I162; I171-I172 I182; J011-J016; J021-J027; J111-J114; J121-J125; J131-J134; J141-J146; J151-J154; J161-J162; J164; J171-J172; J174-J175; J181-J184; J191; J193-J197; J211; J213; J215-J216; J221-J223; J225; J227-J228;
04 Marketing/Sales	G131-G132; G211; G311; G511; G513; G713-G714; G973;
05 Clerical/Administrative	B211; B511-B514; B521-B524; B531-B535; B541-B543; B551-B554; B561-B563; B571-B575; G715; G721; G972;
06 Production Workers	G731-G732; G811; G814; G923-G924; G931-G932; G961-G962; G971; G982-G983; H732-H735; H811-H812; H821-H822; H831-H832; I181; I211-I216; J163; J173; J192; J212; J214; J217; J224: J226; J311-J319;

APPENDIX 3

Editing, Outlier Detection, and Imputation

To maximize the usability of the collected information, one engages in three principal activities, *editing, outlier detection, and imputation*, to ensure that the final data are of the highest quality. Editing is an interactive process whereby the respondent is asked to confirm information that either appears suspect or does not follow some pre-specified general rules governing the data to be collected. This process takes place in the field during data collection.

The detection of outliers is a statistical technique used to identify anomalous responses that either evaded edits, or that did not conform to the correlation structure of the majority of the data (did not follow known relationships). An outlying observation may be classified into two categories, *representative* and *non-representative*. The former has to be left intact as it represents other units in the population that exhibit the same characteristics. The latter, however, should be treated to prevent it from having a significantly positive or negative impact on the estimates. Both types of outliers should be flagged for possible exclusion from imputation.

Imputation is a statistical technique used to fill in information that the respondent fails to provide. It can be applied to records with either partially (certain items have not been collected) or fully (no items have been collected) missing data. This process takes place in the head office after all data have been received and have gone through outlier detection and treatment.

Editing of Data

The workplace questionnaire contains ten distinct blocks. Each block focuses on a different theme. In most cases a single respondent will be able to answer all the questions. If the primary respondent is unable to provide the requested information in its entirety, then he or she will be asked to identify the person privy to this information. The capture vehicle is capable of accepting up to ten different respondents, one for each content block.

The employer CAI (Computer Assisted Interview) capture vehicle performs validity, range, and inter-field edits. These are the types of edits that are performed during the collection of the first wave data. For subsequent waves a suitable set of historical edits has been developed. The majority of inter-field edits are confined to a single content block. If an edit failure occurs between blocks, then the primary respondent is asked to confirm the information.

An example of a validity edit is that total annual expenditures be positive. The corresponding range edit requires that expenditures not exceed an upper bound. A related inter-field edit for total annual expenditures ensures that the sum of annual gross payroll and non-wage expenditures does not exceed total annual expenditures.

The employee CATI (Computer Aided Telephone Interview) application performs validity, range, inter-field and historical edits. Any edit failures are resolved during the telephone interview.

Outlier Detection

The use of CATI for data collection greatly reduces the number of gross response and typographical errors. If either type of error remains undetected, then a *multivariate outlier detection* routine is applied to all complete and partial respondents prior to imputation. The technique uses robust Mahalanobis distance

– a statistic that measures the distance of an observation from the center of the data - to identify units for which this statistic exceeds a pre-specified cut-off defined by a percentile of the corresponding χ^2 distribution. This type of outlier detection is performed for workplaces at the micro data level. The sensitivity of the process can be adjusted to suit the survey's needs.

The current implementation of the outlier detection routine does not incorporate design weights. To be able to use the technique successfully with business survey data, one has to satisfy two criteria: (a) data homogeneity, and (b) data symmetry. Achieving data homogeneity obviates the need to use design weights when pooling neighboring strata to increase the resolution of the outlier routine. Data homogeneity reduces the effect of the design and the complex problem of identifying aberrant observations in a sample drawn from a finite population reduces to a much simpler problem of dealing with outliers in the context of an infinite population.

Homogeneity can be achieved by applying an appropriate function to one or more variables. After data have been suitably transformed (eg., square root, log, etc.), the distributions of the resulting variables should be evaluated for approximate symmetry. This requirement stems from the fact that most outlier detection theory has been developed for contaminated normal distributions. The modified Stahel-Donoho approach is no exception. For WES, approximate symmetry is achieved for ratios of continuous variables to total employment.

The outlier routine can be applied to respondents of a single wave, or across waves. To do so, the response vector $\mathbf{x}_i$ would be modified to include data from two consecutive waves. The possibility of extending the utility of the approach beyond two waves will be studied shortly. Our goal is to develop a method that would fill the gap between cross-sectional outlier detection and robust time series analysis.

Data validation is also performed at a macro level. For a number of key variables we identify the top ten contributors to the weighted estimates for further analysis. Subject matter officers identify both micro and macro level anomalies and correct errors. After errors have been corrected, the data validation cycle is repeated. All remaining outliers are flagged and excluded from imputation.

Imputation

Imputation methods are used cross-sectionally for item non-response for units appearing within each wave for the first time. Longitudinal imputation methods are for wave non-response if historical data are available. In the absence of prior information, total non-response is handled by modifying the weights of the respondents. This approach assumes that the non-response is occurring completely at random.

There are four main imputation methods being used for the first wave of the employer portion of WES: deterministic, distributional, ratio and weighted hot deck. Deterministic imputation is used when a single missing field can be deduced uniquely from the given information. For example, if one component of a sum is missing and the remaining components including the sum are present, then the missing component can be determined uniquely.

Distributional imputation is used for questions where the respondent is asked to provide a total and its breakdown into multiple categories when either two or more of the categories are missing. The distribution of the categories is computed at a macro level and applied at the micro level. To illustrate this approach, let us assume that the respondent gave us total employment but was unable to provide a breakdown by occupational group. We would apply the distribution of the occupational groups computed at the industry/size level to the total employment figure to impute the missing fields.

Ratio imputation is mainly used for continuous variables. The missing value is replaced by the adjusted value of an auxiliary variable from a randomly selected donor within an imputation class. The adjustment usually takes the form of the sum of all donors of the missing variable divided by the sum of the auxiliary variable.

For weighted hot deck, a missing field is imputed using the response of a suitable donor. The donor is selected randomly with a probability of selection equal to the ratio of its sample weight over the sum of the sample weights of all units in the corresponding cross-sectional imputation class. The weighted hot deck approach was adopted for the following four reasons. The method is easy to implement. It leads to approximately *p* unbiased point estimates (Rao, 1996). A consistent variance estimator can be constructed in the presence of imputed data (Rao, 1996). And lastly, most questions are independent keeping the number of post-imputation adjustments to maintain internal data consistency to a minimum.

Missing data on the employee questionnaire are imputed using deterministic and weighted hot-deck imputation. To avoid producing inconsistencies in the data, most interrelated fields are imputed as a block. Since there are a number of questions falling into this category, a post-imputation system has been developed to preserve all inter-field relationships.

APPENDIX 4

Overview of WES Population Estimates

The purpose of this document is to explain in detail the different populations of interest in the Workplace and Employee Survey (WES). This is done to ensure that users of the data are not only aware of the populations which they study, but also, that they are able to relay this message to readers of articles that may produce or estimates they may release. Cautionary notes are given when applicable.

Note: Workplace and location are synonymous in this document. All estimates provided are real estimates from the WES survey. The workplace target population refers to the list of workplaces for which information is desired. The workplace analysis portion refers to the list of workplaces that were sampled and for which data has been made readily available. The employee target population refers to the list of employees for which information is desired. The employee analysis portion refers to the list of employees that were sampled and for which data has been made readily available.

WORKPLACE 1999

Workplace Target Population

The target population for the workplace component is defined as all business locations operating in Canada in March 1999 that have at least one paid employee in March 1999 who receives a Customs Canada and Revenue Agency T-4 Supplementary form, with the following exceptions:

> Workplaces in Yukon, Nunavut and Northwest Territories
> Workplaces operating in crop production and animal production; fishing, hunting and trapping; private households, religious organizations, and public administration.

Workplace Analysis Portion (6322 locations)

The analysis portion is the set of all sampled workplaces that have responded to the 1999 workplace questionnaire, are part of the 1999 workplace target population, and have at least one paid employee in March 1999 who receives a Customs Canada and Revenue Agency T-4 Supplementary form. The analysis portion may be used in conjunction with the weights to reflect the 1999 workplace target population.

Note: The process of re-weighting has been used to account for non-respondent locations, and as a result, the final workplace weights should be used in all analyses. Locations that were sampled but discovered to be out-of-business, out-of-scope, have zero employees, or in receivership in March 1999 are not included in the analysis portion as they are not part of the target population.

Below are a number of examples that use the 1999 workplace analysis portion.

Example 1: Total number of locations in the 1999 workplace target population.

$$\hat{N} = \sum_i w_i = 718{,}083$$

w_i - Final location weight

Example 2: Total number of employees for locations in the 1999 workplace target population.

$$\hat{X} = \sum_i w_i x_i = 10{,}777{,}543$$

w_i - Final location weight
x_i - Employment

Example 3: Average gross payroll per employee in the 1999 workplace target population.

$$\hat{R} = \frac{\sum_i w_i z_i}{\sum_i w_i x_i} = \$31{,}019$$

w_i - Final location weight
x_i - Employment
z_i - Gross payroll

Example 4: Average gross payroll per employee of workplaces that offer non-wage benefits in the 1999 workplace target population.

$$\hat{R}_d = \frac{\sum_i w_i z_i \delta_i}{\sum_i w_i x_i \delta_i} = \$33{,}481$$

w_i - Final location weight
x_i - Employment
z_i - Gross payroll
δ_i - Non-wage benefit indicator (equals 1 if location offers non-wage benefits; 0 otherwise)

EMPLOYEE 1999

Employee Target Population

The target population for the employee component is all employees working or on paid leave in March 1999 who receive a Customs Canada and Revenue Agency T-4 Supplementary form. The aforementioned employee must also belong to a workplace from the 1999 workplace target population.

Employee Analysis Portion (23,540 employees)

The analysis portion is the set of all sampled employees that have responded to the 1999 employee questionnaire, and are part of the 1999 employee target population. The analysis portion may be used in conjunction with the weights to reflect the 1999 employee target population.

Note: The process of re-weighting has been used to account for non-respondent employees, and as a result, the final employee weights should be used in all analyses. Employees that were sampled but discovered to be dead or out-of-scope (not working for the sampled location in March 1999) are not included.

Below are a number of examples that use the 1999 employee analysis portion.

Example 1: Total number of employees in the 1999 employee target population.

$$\hat{N} = \sum_i w_i = 10{,}777{,}543$$

w_i - Final employee weight

Example 2: Average hourly wage per employee in the 1999 employee target population.

$$\hat{\bar{X}} = \frac{\sum_i w_i x_i}{\sum_i w_i} = \$18.53$$

w_i - Final employee weight
x_i - Hourly wage

Example 3: Average hourly wage per employee that is in a union or covered by a collective bargaining agreement (CBA) in the 1999 employee target population.

$$\hat{\bar{X}}_d = \frac{\sum_i w_i x_i \delta_i}{\sum_i w_i \delta_i} = \$20.36$$

w_i - Final employee weight
x_i - Hourly wage
δ_i - Union status indicator (equals 1 if employee is in a union or covered by a CBA; 0 otherwise)

LINKED WORKPLACE/EMPLOYEE 1999

Linked Target Population

The 1999 linked target population is the set of locations from the 1999 workplace target population and employees from the 1999 employee target population.

Linked Analysis Portion (5,733 locations; 23,540 employees)

The linked analysis portion consists of workplaces from the 1999 workplace analysis portion with at least one responding employee and employees from the 1999 employee analysis portion. The analysis portion may be used in conjunction with the weights to reflect the 1999 linked target population.

Note: When performing employee analysis, linking to workplace characteristics, one should use the employee final weights, in association with the complete employee file. When performing workplace analysis, linking to employee characteristics, the workplace linked weight should be used considering only workplaces with at least one responding employee. Re-weighting is performed to adjust for workplaces with no-responding employees.

Example 1: Average hourly wage per employee working for a non-profit workplace in the 1999 linked target population.

$$\hat{\bar{X}}_d = \frac{\sum_i w_i x_i \delta_i}{\sum_i w_i \delta_i} = \$21.51$$

w_i - Final employee weight
x_i - Hourly wage
δ_i - Non-profit indicator (from location file; equals 1 if location is a non-profit workplace; 0 otherwise)

Example 2: Average hourly wage per employee that is in a union or covered by a collective bargaining agreement and working for a non-profit workplace in the 1999 linked target population.

$$\hat{\bar{X}}_d = \frac{\sum_i w_i x_i \delta_{1i} \delta_{2i}}{\sum_i w_i \delta_{1i} \delta_{2i}} = \$22.05$$

w_i - Final employee weight
x_i - Hourly wage
δ_{1i} - Union status indicator (equals 1 if employee is in a union or covered by a CBA; 0 otherwise)
δ_{2i} - Non-profit indicator (from location file; equals 1 if location is a non-profit workplace; 0 otherwise)

Example 3: Average gross payroll per employee for workplaces with at least one employee with a long-term disability in the 1999 linked target population.

$$\hat{R}_d = \frac{\sum_i w_i z_i \delta_i}{\sum_i w_i x_i \delta_i} = \$32{,}810$$

w_i - Linked location weight
x_i - Employment
z_i - Gross Payroll
δ_i - Long-term disability indicator (from employee file; equals 1 if location has at least one employee with a long-term disability; 0 otherwise)

Example 4: Average gross payroll per employee of locations that offer non-wage benefits in the 1999 linked target population with at least one employee with a long-term disability.

$$\hat{R}_d = \frac{\sum_i w_i z_i \delta_{1i} \delta_{2i}}{\sum_i w_i x_i \delta_{1i} \delta_{2i}} = \$32{,}790$$

w_i - Linked location weight
x_i - Employment
z_i - Gross Payroll
δ_{1i} - Non-wage benefit indicator (equals 1 if location offers non-wage benefits; 0 otherwise)
δ_{2i} - Long-term disability indicator (from employee file; equals 1 if location has at least one employee with a long-term disability; 0 otherwise)

WORKPLACE 2000

Workplace Target Population

The WES is a longitudinal survey with its workplace component being refreshed every second year (2001, 2003, etc.). For this reason, the 2000 workplace target population remains unchanged from 1999.

Workplace Analysis Portion (6,068 locations)

The 2000 analysis portion is the subset of workplaces from the 1999 workplace analysis portion, having at least one paid employee in March 2000 who receives a Customs Canada and Revenue Agency T-4 Supplementary form. Excluded (considered out-of-scope) from the 2000 workplace analysis portion are workplaces that in March 2000:

> Are located in the Yukon, Nunavut or Northwest Territories
> Are operating in crop production and animal production; fishing, hunting and trapping; private households, religious organizations, and public administration.

These exclusions only apply to the analysis portion of 2000 and not the target population.

Note: The final workplace weights should be used in the analyses as re-weighting has been performed to account for non-respondents from 1999. Analyses performed on the 2000 workplace analysis portion do not represent the cross-sectional picture of all workplaces in March 2000. This stems from the fact that workplaces which came into existence after the creation of the 1999 frame have a zero probability of being included in the sample and no re-weighting has been done to account for them. Thus, all analyses from the 2000 workplace analysis portion should refer to continuing (still in-business and in-scope) units from the 1999 population only.

Below are a number of examples that use the 2000 workplace analysis portion.

Example 1: Total number of continuing locations in the 2000 workplace target population.

$$\hat{N} = \sum_i w_i = 668{,}188$$

w_i - Final location weight

Example 2: Total number of employees in continuing locations in the 2000 WES workplace target population.

$$\hat{X} = \sum_i w_i x_i = 10{,}785{,}150$$

w_i - Final location weight
x_i - Employment

Example 3: Average gross payroll per employee of continuing locations in the 2000 workplace target population.

$$\hat{R} = \frac{\sum_i w_i z_i}{\sum_i w_i x_i} = \$32{,}166$$

w_i - Final location weight
x_i - Employment
z_i - Gross payroll

Example 4: Average gross payroll per employee of continuing locations that offer non-wage benefits in the 2000 workplace target population.

$$\hat{R}_d = \frac{\sum_i w_i z_i \delta_i}{\sum_i w_i x_i \delta_i} = \$34{,}988$$

w_i - Final location weight
x_i - Employment
z_i - Gross payroll
δ_i - Non-wage benefit indicator (equals 1 if location offers non-wage benefits; 0 otherwise)

EMPLOYEE 2000

Employee Target Population

The WES is a longitudinal survey with its employee component being refreshed every second year (2001, 2003, etc.). For this reason, the 2000 employee target population remains unchanged from 1999.

Employee Analysis Portion (20,167 employees)

The 2000 analysis portion is the subset of employees from the 1999 employee analysis portion whose employer of March 1999 is part of the 2000 workplace analysis portion. This set of employees is split between continuers (working for same employer in March 1999 and March 2000) and exiters (no longer working for the same employer as March 1999). The set of exiters either works for a new employer that may or may not be part of the 2000 workplace target population or is no longer in the workforce.

Excluded from the 2000 employee analysis portion are employees that belong to locations that are excluded from the 2000 workplace analysis portion. These exclusions only apply to the analysis portion of 2000 and not the target population.

Note: The final employee weights should be used in the analyses as re-weighting has been performed to account for 1999 and 2000 non-respondents. Analyses performed on the 2000 employee analysis portion do not correspond to all employees as of March 2000. This stems from the fact that employees belonging to workplaces which came into existence after the creation of the 1999 frame have a zero probability of being included in the sample and no re-weighting has been done to account for them. Thus, all analyses from the 2000 employee analysis portion should refer to continuing or exiting units from the 1999 population only.

Below are a number of examples that use the 2000 employee analysis portion.

Example 1: Total number of continuing or exiting employees in March 2000 working in March 1999 for a continuing workplace. (ie. Employee belonged in March 1999 to a workplace that is part of the 2000 analysis portion)

$$\hat{N} = \sum_i w_i = 10{,}755{,}029$$

w_i - Final employee weight

Example 2: Total number of continuing employees in March 2000 working in March 1999 and March 2000 for the same continuing workplace.

$$\hat{N}_d = \sum_i w_i \delta_i = 8{,}964{,}798$$

w_i - Final employee weight
δ_i - Continuer status indicator (equals 1 if employee is working for the same employer in March 2000 as in March 1999; 0 otherwise)

Example 3: Total number of exiting employees between April 1999 and March 2000 working in March 1999 for a continuing workplace.

$$\hat{N}_d = \sum_i w_i \delta_i = 1{,}790{,}230$$

w_i - Final employee weight
δ_i - Exiter status indicator (equals 1 if employee is, in March 2000, no longer working for the same employer as in March 1999; 0 otherwise)

Example 4: Average hourly wage per working employee in March 2000 working in March 1999 for a continuing workplace.

$$\hat{\bar{X}}_d = \frac{\sum_i w_i x_i \delta_i}{\sum_i w_i \delta_i} = \$18.11$$

w_i - Final employee weight
x_i - Hourly wage
δ_i - Working status indicator (equals 1 if employee is working)

LINKED ANALYSIS OF WORKPLACE AND EMPLOYEE 2000

Linked Target Population

The 2000 linked target population is the set of locations from the 2000 workplace target population and employees from the 2000 employee target population.

Linked Analysis Portion (5,453 locations; 20,167 employees)

The linked analysis portion consists of workplaces from the 2000 workplace analysis portion with at least one responding employee and employees from the 2000 employee analysis portion. The analysis portion may be used in conjunction with the weights to reflect the 2000 linked target population.

Note: When performing employee analysis, linking to workplace characteristics, one should use the employee final weights, in association with the complete employee file. When performing workplace analysis, linking to employee characteristics, the workplace linked weight should be used, considering only workplaces with at least one responding employee. Re-weighting is performed to adjust for workplaces with no-responding employees. Analyses performed on the 2000 linked analysis portion do not represent the cross-sectional picture of all linked workplace/employees in March 2000. This stems from the fact that workplaces and employees belonging to workplaces which came into existence after the creation of the 1999 frame have a zero probability of being included in the sample and no re-weighting has been done to account for them. Thus, all analyses from the 2000 linked analysis portion should refer to continuing or exiting employees from continuing locations.

Below are a number of examples that use the 2000 linked analysis portion.

Example 1: Average hourly wage per employee who in March 1999 were working in a continuing workplace that during the 2000 collection, was a non-profit workplace. The employee may or may not still work for the same employer as in March 1999.

$$\hat{\bar{X}}_d = \frac{\sum_i w_i x_i \delta_i}{\sum_i w_i \delta_i} = \$20.90$$

w_i - Final employee weight
x_i - Hourly wage
δ_i - Non-profit indicator (from location file; equals 1 if location is a non-profit workplace; 0 otherwise)

Example 2: Average hourly wage per employee who is working for a non-profit workplace in the 2000 linked target population and was working for the same location as March 1999.

$$\hat{\bar{X}}_d = \frac{\sum_i w_i x_i \delta_{1i} \delta_{2i}}{\sum_i w_i \delta_{1i} \delta_{2i}} = \$22.79$$

w_i - Final employee weight
x_i - Hourly wage
δ_{1i} - Continuer status indicator (equals 1 if employee is working in for the same employer in March 2000 as in March 1999; 0 otherwise)
δ_{2i} - Non-profit indicator (from location file; equals 1 if location is a non-profit workplace; 0 otherwise)

Example 3: Average gross payroll per employee for continuing workplaces with at least one continuing or exiting employee with a long-term disability in March 2000.

$$\hat{R}_d = \frac{\sum_i w_i z_i \delta_i}{\sum_i w_i x_i \delta_i} = \$34{,}619$$

w_i - Linked location weight
x_i - Employment
z_i - Gross Payroll
δ_i - Long-term disability indicator (from employee file; equals 1 if location has at least one employee with a long-term disability; 0 otherwise)

Example 4: Average gross payroll per employee for continuing workplaces with at least one exiting employee with a long-term disability in March 2000.

$$\hat{R}_d = \frac{\sum_i w_i z_i \delta_i}{\sum_i w_i x_i \delta_i} = \$34{,}264$$

w_i - Linked location weight
x_i - Employment
z_i - Gross Payroll
δ_i - Long-term disability, exiter indicator (from employee file; equals 1 if location has at least one exiting employee with a long-term disability; 0 otherwise)

LONGITUDINAL WORKPLACE 1999/2000

Longitudinal Workplace Target Population

The longitudinal workplace target population is the same as the 2000 workplace target population.

Longitudinal Workplace Analysis Portion (6,068 locations)

The longitudinal workplace analysis portion is the same as the 2000 workplace analysis portion including data from both 1999 and 2000.

Note: Longitudinal estimates calculated from 1999 in the following examples are done so using only continuing locations.

Below are a number of examples that use the longitudinal workplace analysis portion.

Example 1: Percentage change in total revenue from 1999 to 2000 for continuing locations.

$$\hat{P} = \frac{\sum_i w_i x_{i2000} - \sum_i w_i x_{i1999}}{\sum_i w_i x_{i1999}} \times 100 = 6.95\%$$

w_i - Final location weight
x_{i2000} - 2000 Revenue
x_{i1999} - 1999 Revenue

Example 2: Percentage change in average gross payroll per employee from 1999 to 2000 for continuing locations.

$$\hat{P} = \frac{\frac{\sum_i w_i z_{i2000}}{\sum_i w_i x_{i2000}} - \frac{\sum_i w_i z_{i1999}}{\sum_i w_i x_{i1999}}}{\frac{\sum_i w_i z_{i1999}}{\sum_i w_i x_{i1999}}} \times 100 = 3.23\%$$

w_i - Final location weight
x_{i2000} - 2000 Employment
x_{i1999} - 1999 Employment
z_{i2000} - 2000 Gross Payroll
z_{i1999} - 1999 Gross Payroll

Example 3: Percentage change in total revenue for locations offering non-wage benefits in both years for continuing locations.

$$\hat{P} = \frac{\sum_i w_i x_{i2000} \delta_i - \sum_i w_i x_{i1999} \delta_i}{\sum_i w_i x_{i1999} \delta_i} \times 100 = 6.21\%$$

w_i - Final location weight
x_{i2000} - 2000 Revenue
x_{i1999} - 1999 Revenue
δ_i - Non-wage benefit indicator (equals 1 if location offers non-wage benefits in both survey years, 1999 and 2000; 0 otherwise)

LONGITUDINAL EMPLOYEE 1999/2000

Longitudinal Employee Target Population

The longitudinal employee target population is the same as the 2000 employee target population.

Longitudinal Employee Analysis Portion (20,167 employees)

The longitudinal employee analysis portion is the same as the 2000 employee analysis portion including data from both 1999 and 2000.

Note: For longitudinal analyses the 2000 employee final weights should be used. Longitudinal estimates calculated from 1999 in the following examples are done so using only employees who in March 1999 were part of continuing locations.

Below are a number of examples that use the longitudinal employee analysis portion.

Example 1: Percentage change in average hourly wage per employee between 1999 and 2000 working in March 1999 for a continuing location. (Employee may be working for the same location as in March 1999, working for a new location, or not working at all.)

$$\hat{P} = \frac{\frac{\sum_i w_i x_{i2000}}{\sum_i w_i} - \frac{\sum_i w_i x_{i1999}}{\sum_i w_i}}{\frac{\sum_i w_i x_{i1999}}{\sum_i w_i}} \times 100 = -2.64\%$$

w_i - Final 2000 employee weight
x_{i2000} - 2000 Hourly Wage
x_{i1999} - 1999 Hourly Wage

Example 2: Percentage change in average hourly wage per continuing employee between 1999 and 2000 working in March 1999 for a continuing location.

$$\hat{P} = \frac{\frac{\sum_i w_i x_{i2000}\delta_i}{\sum_i w_i\delta_i} - \frac{\sum_i w_i x_{i1999}\delta_i}{\sum_i w_i\delta_i}}{\frac{\sum_i w_i x_{i1999}\delta_i}{\sum_i w_i\delta_i}} \times 100 = 3.40\%$$

w_i - Final location weight
x_{i2000} - 2000 Hourly Wage
x_{i1999} - 1999 Hourly Wage
δ_i - Continuer status indicator (equals 1 if employee is working in for the same employer in March 2000 as in March 1999; 0 otherwise)

Example 3: Percentage change in average hourly wage per exiting employee between 1999 and 2000 working in March 1999 for a continuing location and working in March 2000 for a new employer.

$$\hat{P} = \frac{\frac{\sum_i w_i x_{i2000} \delta_i}{\sum_i w_i \delta_i} - \frac{\sum_i w_i x_{i1999} \delta_i}{\sum_i w_i \delta_i}}{\frac{\sum_i w_i x_{i1999} \delta_i}{\sum_i w_i \delta_i}} \times 100 = 4.74\%$$

w_i - Final location weight
x_{i2000} - 2000 Hourly Wage
x_{i1999} - 1999 Hourly Wage
δ_i - Exiter status indicator (equals 1 if employee is, in March 2000, no longer working for the same employer as in March 1999; 0 otherwise)

LONGITUDINAL LINKED WORKPLACE/EMPLOYEE 1999/2000

Longitudinal Linked Target Population

The longitudinal linked target population is the same as the 2000 linked target population.

Longitudinal Linked Analysis Portion (5,453 locations; 20,167 employees)

The longitudinal linked analysis portion is the same as the 2000 linked analysis portion including data from both 1999 and 2000.

Note: When performing longitudinal employee analysis, linking to workplace characteristics, one should use the 2000 employee final weights, in association with the complete employee file. When performing longitudinal workplace analysis, linking to employee characteristics, the 2000 workplace linked weight should be used, considering only workplaces with at least one responding employee. Re-weighting is performed to adjust for workplaces with no responding employees. Longitudinal estimates calculated from 1999 in the following examples are done so using only employees who in March 1999 were part of continuing locations, regardless of where they work (or don't work) in March 2000. Also included in the examples are the continuing locations.

Below are a number of examples that use the longitudinal linked analysis portion.

Example 1: Percentage change in average hourly wage per employee who in March 1999 was working for a non-profit continuing workplace. (Employee may be working for the same location as in March 1999, working for a new location, or not working at all.)

$$\hat{P} = \frac{\frac{\sum_i w_i x_{i2000} \delta_i}{\sum_i w_i \delta_i} - \frac{\sum_i w_i x_{i1999} \delta_i}{\sum_i w_i \delta_i}}{\frac{\sum_i w_i x_{i1999} \delta_i}{\sum_i w_i \delta_i}} \times 100 = -2.45\%$$

w_i - Final employee weight
x_{i2000} - 2000 Hourly Wage
x_{i1999} - 1999 Hourly Wage
δ_i - Non-profit indicator (from location file; equals 1 if location was a non-profit workplace in 1999; 0 otherwise)

Example 2: Percentage change in average hourly wage per continuing employee who in March 1999 was working for a continuing location. The location was non-profit in 1999 and 2000.

$$\hat{P} = \frac{\frac{\sum_i w_i x_{i2000} \delta_{1i} \delta_{2i}}{\sum_i w_i \delta_{1i} \delta_{2i}} - \frac{\sum_i w_i x_{i1999} \delta_{1i} \delta_{2i}}{\sum_i w_i \delta_{1i} \delta_{2i}}}{\frac{\sum_i w_i x_{i1999} \delta_{1i} \delta_{2i}}{\sum_i w_i \delta_{1i} \delta_{2i}}} \times 100 = 4.87\%$$

w_i - Final employee weight
x_{i2000} - 2000 Hourly Wage
x_{i1999} - 1999 Hourly Wage
δ_{1i} - Continuer status indicator (equals 1 if employee is working in for the same employer in March 2000 as in March 1999; 0 otherwise)
δ_{2i} - Non-profit indicator (from location file; equals 1 if location is a non-profit workplace in 1999 and 2000; 0 otherwise)

Example 3: Percentage change in total revenue from 1999 to 2000 for continuing workplaces with at least one continuing or exiting employee with a long-term disability in March 1999 and March 2000 in the longitudinal linked target population.

$$\hat{P} = \frac{\sum_i w_i x_{i2000} \delta_i - \sum_i w_i x_{i1999} \delta_i}{\sum_i w_i x_{i1999} \delta_i} \times 100 = 2.93\%$$

w_i - Final linked location weight
x_{i2000} - 2000 Revenue
x_{i1999} - 1999 Revenue
δ_i - Long-term disability indicator (equals 1 if employee has long-term disability in 1999 and 2000; 0 otherwise)

APPENDIX 5

Linked Analysis

Why linked models must be treated differently

With linked employer and employee data such as Statistics Canada's Workplace and Employee Survey (WES), researchers are provided an opportunity to investigate business and labour market outcomes that depend critically on the interactions between employers and employees. At the same time, they will also have to face some statistical and econometric problems in their modelling of the business and labour market activities.

Since the late 1990s, economists have proposed a variety of empirical models that can be estimated with linked (matched) employer-employee data.[1] Although the models employed by these studies are basically the familiar linear regression function, there are a number of new elements embedded in these models warranting a treatment different from the classical linear regression analysis. Consider a linear model specified for some employee-level outcome Y_{ij} in which employee i is characterised by X_{ij} and establishment j is characterised by Z_j:

$$Y_{ij} = \alpha_j + \beta_j X_{ij} + \varepsilon_{ij},$$
$$\alpha_j = \alpha_0 + \alpha_1 Z_j + u_j,$$
$$\beta_j = \beta_0 + \beta_1 Z_j + v_j,$$

where ε_{ij}, u_j, and v_j are classical disturbances, ε_{ij} is independent of X_{ij}, u_j and v_j are independent from each other and they are independent of Z_j. A linear model can be derived from these specifications:

$$Y_{ij} = \alpha_0 + \alpha_1 Z_j + \beta_0 X_{ij} + \beta_1 X_{ij} Z_j + u_j + v_j X_{ij} + \varepsilon_{ij}.$$

Models like the above, often referred to as *mixed models* (*varying parameter models*), contains stochastic elements (u_j and v_j) that are not observable to the analyst. Classical linear regression analysis applies to the above model only if $u_j = v_j = 0$. When $v_j = 0$, it becomes an example of the *error component models*, and when $u_j = 0$, we obtain an example of the *random coefficients models*.

The mixed model becomes more complex if we attempt to analyse outcomes of the interactions between employers and employees over time. Even in the absence of error components and random coefficients, some of the standard assumptions of the classical regression analysis are quite likely to be violated in a mixed model. In particular, intra-firm correlation, inter-firm heteroscedasticity, measurement error brought by aggregation can all cause serious consequences if these problems are not carefully addressed. Furthermore, the full model, capable of capturing the effects of employer and employee characteristics and the effects of decisions (choices) made by employers and employees, is not necessarily hierarchical or balanced[2]. Hence, not all the treatments established by the multilevel modelling literature[3] are applicable in such a specification.

[1] See Abowd and Kramarz (1999) for a review. Haltiwanger et al. (1999) eds. contains selected articles presented at the 1998 international Symposium of Linked Employer-Employee Data .
[2] The basic linear model employed by Abowd and Kramarz (1999) for their review is such an example.
[3] See Goldstein (1995) for an introduction to multilevel analysis.

Using employer variables in employee analyses

When one attempts to analyze employee level outcomes using variables at the employer level, a disaggregation of the employer variables is initiated. Employees drawn from the same firm or establishment would have identical employer variables such as technology investment, training expenditure and industry, and these employer variables may not be independent across workers within the same establishment. But parameter estimation necessarily treats the value of an employer variable associated with each employee within the same establishment as independent information. As a result, some estimates may be spuriously different from 0. In order to avoid this, one shall need to correct the downward bias in the estimated standard errors. The correction procedure is discussed in Moulton (1985) and Troske (1996).

One may follow the classical regression analysis to assume homogeneous employees within a firm, but it is likely that employees between firms are heterogeneous. Wrong inference can be made if grouped data drawn from a heterogeneous population are treated as if they are drawn from a homogeneous one. The group-wise heteroscedasticity problem, however, is not a new issue. Treatments are discussed in many standard econometric textbooks[4]. A random coefficients model specification, due to Hildreth and Houck (1968), might be a convenient way out of the problem.

Using summarized employee data in employer analyses

Information collected from employees could be of particular interest for researchers modelling employer outcomes. But many variables defined at the employee level might be problematic when being used at the employer level, particularly those based on the subjective assessments made by the surveyed employees. Hence, in linked analyses, the *error in variables* problem brought by aggregation becomes a norm rather than an exception.

The solution to measurement error is to replace the variable in question by an instrumental variable (IV), a variable that is highly correlated with the true value of the underlying variable but not with the measurement error[5]. The IV estimators are asymptotically consistent, efficient, and normal under certain general conditions. Fuller (1987) is an excellent reference on the IV method. A suitable instrument is not easy to find in many situations, but linked data makes it easier for analysts to find good instruments. However, correcting problems induced by measurement error is not the only usefulness of the IV method. More importantly, the IV method is employed by many empirical studies to solve the possible endogeneity problem: an explanatory variable in a model depends also on the dependent variable. In the classical regression context, this is the case where the explanatory variable is correlated with the error term. The endogeneity problem makes the IV method (in stead of the multilevel model) more popular in linked employer-employee analysis.

Software

The mixed model estimation, the IV method and estimations of fixed effects, random effects can be handled by many statistical/econometric programs. SAS and STATA are two powerful packages. In SAS, the GLM and the MIXED PROCs can be used for estimation of the multilevel model, taking weights into

[4] See for example, chapters 16 and 17 of Judge et al (1982).
[5] The measurement error can be non-classical in the sense that it is not independent of the true values of the variable in question. See Barron, Berger and Black (1999).

the procedures. STATA can offer capacities to estimate many models researchers may specify and provide a number of procedures that account for complex sample design effects (with the “svy” prefix). However, users should be aware that STATA is not able to correctly incorporate the dead units. If the domain of interest is used, the point estimates will be correct but not their variances. By far, the use of bootstrap weights in SAS regression procedures is the most general and practical way to generating design-based estimates and variances.

The WES project team is testing a number of other software packages appropriate to mixed models in 2002.

APPENDIX 6

Weighting and Estimation

The Workplace and Employee Survey is a sample of Canadian business locations from which a certain number of employees is selected depending on the size of the location measured by total employment.

Weighting

Having selected a sample from the population of interest means that analysis variables will be collected for only a fraction of available units. To be able to produce estimates that relate to the population, each sampled unit is assigned a weight to represent other, similar units that have not been selected. This weight, called the *design weight*, is equal to the inverse of the probability of selection. For example, if one selects two units from among ten, then each unit is given a weight of five.

The WES employer sample is selected independently in 252 strata without replacement (a unit is not replaced in the stratum after it has been drawn). The strata represent homogeneous groups of units identified by industry (14 classes), region (6 classes), and size (3 classes). Given an overall permissible sample size, the number of units selected in each stratum is computed such that no one cell exceeds a pre-specified coefficient of variation. This means that highly variable strata are sampled at a higher rate and vice-versa.

The initial sample determines the design weight of each unit. Throughout the survey process the initial design weights may undergo several adjustments, which strive to protect the representativity of the sample. For WES two adjustments were made, one to compensate for complete non-response and one to diminish the influence of stratum jumpers (large units believed to be small and vice-versa) on estimates. To adjust for non-response one multiplies the weights of responding units by a ratio of all sampled units to all responding units within each stratum. This process is predicated on the assumption that respondents and non-respondents behave alike. Since non-response exists mainly amongst the smaller units, this assumption is not unreasonable.

Adjusting for stratum jumpers is more complex as there are at least three methods for dealing with the problem. One can either decrease the design weight of the stratum jumper and distribute the difference over the remaining units within the stratum, or one can reduce its values, or one can remove the unit entirely and treat it as non-response. We selected the first option where we targeted approximately 30 employers for a design weight adjustment.

The use of the design weights, whether initial or final, results in an unbiased yet inefficient estimate. To improve the efficiency of the estimation process, one can benchmark, or *calibrate*, the sample to a set of known or efficiently estimated population totals. In WES this is done using total employment as estimated by SEPH (Survey of Employees, Payroll, and Hours) at the industry by region level, at which the WES estimates are forced to agree with the SEPH estimates. The resulting adjustment factors are applied to the final design weights. Benchmarking is of the most benefit in situations where the calibration variable is highly correlated with the variables of interest.

The product of the final design weight and the calibration factor is the final unit weight. It is used for computing first order statistics such as totals, means, regression coefficients, etc. To calculate second order statistics, or variances, one has to use software packages that allow the user to specify the survey

design. If one uses products such as SAS without suitably transforming the survey weights, the resulting underestimation of the variance may be quite severe.

Estimation

There are many avenues open to the analysts wishing to produce design consistent variances. One is to use the Statistics Canada Generalized Estimations System (GES) that will handle the estimation of totals, means, and ratios for a variety of designs. The use of GES by external researchers may be financially prohibitive given its licensing structure.

A second option is by far the most general and the easiest to put into practice. It involves the use of *bootstrap* weights. Bootstrap is a statistical technique whereby one uses a resampling procedure to generate a number of sets of weights that, if used correctly, capture the variability of a wide variety of statistics. The idea is to compute a large number of bootstrap estimates and then calculate their variance.

Once the bootstrap weights are computed, they can be specified in the weight statement in any SAS procedure that has one. To calculate the variance for a desired variable, one has to produce an estimate based on each set of bootstrap weights. Then one computes the variance of the estimates so produced and applies an adjustment to make the variance design consistent. For examples on how to use the bootstrap method refer to Appendix 7.

APPENDIX 7

Variance Calculation

The use of Bootstrap weights for computing design consistent variances

When one computes the variances for estimates based on samples coming from finite populations, one has to account for the design. This is not easily done in most statistical analysis software packages. Although most of them do allow the use of weights, they do not use them in the proper manner thus resulting in the underestimation of the variance. This could have dire consequences for hypothesis testing and for constructing confidence intervals.

Over the years statistical agencies have developed systems to deal with finite populations but most of them lack the flexibility needed to do data analysis. This is where BOOTSTRAP comes in. It is a technique based on re-sampling. One uses the original sample, from which one selects a simple random sample with replacement of as many units as one has at the outset. This procedure is repeated many times to guarantee consistency.

Once the bootstrap weights are computed, they can be specified in the weight statement in any SAS procedure that has one. To calculate the variance for a desired variable, one has to produce an estimate based on each set of bootstrap weights. Then one computes the variance of the estimates so produced and applies an adjustment to make the variance design consistent. Below are two examples of how this can be achieved for totals and for correlation coefficients (Note: the correlation coefficient is a complex statistic and there is not a method that computes its variance exactly. The best methods achieve 85% to 90% coverage probability for the 95% nominal coverage level.).

Depending on your analysis you would use either the wkp_bsw1-wkp_bsw100, emp_bsw1-emp_bsw1100 or lnk_bsw1-lnk_bsw100. The following example looks at workplace information.

```
PROC SUMMARY DATA = WES NWAY;
        CLASS DOM_IND;
        VAR WKP_FINAL_WT WKP_BSW1-WKP_BSW100;
        WEIGHT TTL_EMP;
        OUTPUT OUT = ESTIM (DROP = _FREQ_ _TYPE_)
                    SUM = EMPL WKP_BSW1-WKP_BSW100;
RUN;

PROC TRANSPOSE DATA = ESTIM
                      OUT = T_ESTIM (DROP = _NAME_ RENAME = (COL1 = ESTIM));
        VAR WKP_BSW1-WKP_BSW100;
                BY DOM_IND;
RUN;

PROC SUMMARY DATA = T_ESTIM NWAY;
        CLASS DOM_IND;
        VAR ESTIM;
        OUTPUT OUT = VAR (DROP = _FREQ_ _TYPE_)
                    CSS = VAR;
RUN;

DATA ESTIM;
        MERGE ESTIM (KEEP = DOM_IND EMPL)
                  VAR;
                BY DOM_IND;
        CV = ROUND (SQRT(50 / 100 * VAR) / EMPL, 0.01);
RUN;
```

The first SUMMARY procedure uses a trick that allows one to compute all necessary estimates in one simple step. This can only be done when one is producing estimates for a single variable. The trick is to specify the bootstrap weights as the analysis variables and to use the analysis variable as the weight. The estimates are computed at the domain industry level specified by the class statement.

After estimates have been computed, transposed and renamed, another SUMMARY procedure is used to compute their variance (actually, their corrected sum of squares, or CSS in SAS). And finally, multiplying the CSS by 50 / 100 produces the correct design based variance. The denominator (100) is the normal adjustment *n* that yields the classical variance. The numerator (50) reflects the fact that each set of bootstrap weights has been averaged over 50 iterations, resulting in an average bootstrap weight. Therefore, the adjustment injects back the variability that has been lost by using the average.

The next example illustrates the use of bootstrap weights for computing correlation coefficients. Here, one has to use a macro to compute individual coefficients, as one cannot easily use the above trick.

```
%MACRO COR_COEF;
        %DO I = 1 %TO 100;
                PROC CORR DATA = BOOT OUTP = CORRS NOPRINT;
                        VAR TTL_EMP CBA_EMP;
                                BY DOM_IND;
                        WEIGHT WKP_BSW&I;
                RUN;

                DATA CORRS (KEEP = DOM_IND CBA_EMP RENAME = (CBA_EMP = CORR));
                        SET CORRS (WHERE = (_TYPE_ = 'CORR' & _NAME_ = 'TTL_EMP'));
                RUN;

                PROC DATASETS FORCE NOLIST;
                        APPEND BASE = ESTIM DATA = CORRS;
                        QUIT;
                RUN;
        %END;
%MEND;

%COR_COEF;

PROC SUMMARY DATA = ESTIM NWAY;
        CLASS DOM_IND;
        VAR CORR;
        OUTPUT OUT = VAR (DROP = _FREQ_ _TYPE_)
                CSS = VAR;
RUN;

PROC CORR DATA = BOOT OUTP = CORRS NOPRINT;
        VAR TTL_EMP CBA_EMP;
                BY DOM_IND;
        WEIGHT WKP_FINAL_WT;
RUN;

DATA CORRS (KEEP = DOM_IND CBA_EMP RENAME = (CBA_EMP = EST_CORR));
        SET CORRS (WHERE = (_TYPE_ = 'CORR' & _NAME_ = 'TTL_EMP'));
RUN;

DATA ESTIM;
        MERGE VAR CORRS;
                BY DOM_IND;
        CV = ROUND(SQRT(50 / 100 * VAR) / EST_CORR * 100, 0.01);
RUN;
```

The macro COR_COEF computes correlation coefficients based on each set of bootstrap weights. The example here treats two continuous variables but may be easily extended to multiple variables both continuous and categorical. After estimates have been computed, the corrected sum of squares is produced along with a correlation coefficient that is based on the final weights.

The two files are then merged, the corrected sum of squares is adjusted and a CV is computed. Similar steps should be followed for computing variances of regression estimates, principal components, and other statistic. With the exception of totals of a single variable the computations cannot be done in one step. To reduce computing time per iteration it is recommended that the initial data set be reduced to the analysis variables.

APPENDIX 8

Deemed Employee Access to Workplace and Employee Survey Microdata

Researchers under agreement with Statistics Canada

A8.1 Steps to follow for entry of Statistics Canada

1. Researchers are to submit proposals to Statistics Canada (STC). Be sure to include in your proposal a justification for using STC microdata. Guidelines and forms can be obtained from your STC analyst.
2. Statistics Canada will carry out a review of the proposal and will notify the primary researcher of the final decision made by the review committee. Ideally this will happen within two months of the date of submission. At that time, Statistics Canada will conduct a security check on all researchers who will be accessing the data. Note that all proposal decisions can be appealed through Statistics Canada.
3. Researchers should contact the STC analyst to indicate their intent before they would like to access the data. Upon that contact, four things will happen:

    - The primary researcher will sign a memorandum of understanding between the project team members and Statistics Canada.
    - The researchers will attend an orientation session (approximately three hours) conducted by the STC analyst.
    - At the end of the session, the STC analyst will administer the oath of office.
    - Researchers who have signed the oath of office will then receive their own pass to access the STC area.

4. Researchers are asked to sign up for a workstation on the days they would like to access data.

5. Data access begins.

A8.2 Steps to follow for submission of output for disclosure analysis

Note: We encourage you to request of STC only the output that is essential to your report . The more requests on which the STC analysts have to perform disclosure analysis, the more difficult it becomes to address all researchers' needs in a timely fashion.

Please follow these steps if you would like to remove output from the STC:

1. Create a subdirectory under your assigned directory containing the files you would like to remove and accompanying analysis that may be necessary for disclosure analysis.
2. Schedule time with the STC analyst to discuss the disclosure analysis. Depending on the level of difficulty of the analysis and the volume of output, the STC analyst may request your presence during the disclosure analysis.
3. Revise your output based on the recommendations of the STC analyst and rename your files under the same subdirectories. Note that additional sessions may be required until all issues are addressed.
4. Advise the STC analyst that the revisions have been made and provide a diskette to transfer the output or indicate that you would like a printed copy.
5. Pick up your copy/diskette from the STC analyst.

NO SURVEY DATA SHOULD BE REMOVED FROM STATISTICS CANADA OR THE RESEARCH DATA CENTRES!

A8.3 Steps to follow to gain access to a database not requested in the original proposal

Normally Statistics Canada will not allow researchers access to a new database if it was not requested in the original proposal. However, this need may arise from time to time. Talk with your STC analyst to determine whether your request can be fulfilled.

1. Researchers must submit a short written request to the STC analyst outlining the rationale for gaining access to a new database in order to achieve the goals of the original proposal.
2. The STC analyst will review your request with Statistics Canada staff, who may ask you for details.
3. If Statistics Canada approves the request, the STC analyst will arrange access to this database.

Note: Unsuccessful applicants are encouraged to submit a new proposal to gain access to additional databases.

A8.4 Steps to follow to add/remove a new researcher to/from a project after acceptance of a proposal by Statistics Canada

Note: Primary researchers are required to include the names of all researchers who are associated with the proposal, particularly any research assistants who will be accessing data in the STC area. However, an occasion may arise when a research assistant may be substituted or added.

Adding a researcher to a project:

1. Primary researchers should indicate to the STC analyst, in writing, the names of researchers who are to be added to the data access for a particular project.
2. The STC analyst will send the primary researcher the appropriate forms to be completed for the security check.
3. The STC analyst will inform the primary researcher of the results of the security check.
4. If the results are acceptable, then the new researcher can contact the STC analyst to arrange a time to attend an orientation session, take the oath of office and receive a security key and password.

Removing a researcher from a project:

1. Primary researchers should indicate to the STC analyst, in writing, the names of any researchers who will no longer be accessing data under this project. The primary researcher should also indicate if the computer files of this researcher should be retained, purged, or reassigned.
2. These researchers will be asked to return their security passes to the STC analyst.

Note: The oath of office remains in effect for these researchers.

A8.5 Steps to follow to exit the STC upon completion of a project

1. Researchers are to submit a draft of the Statistics Canada product to the STC analyst under the conditions of the memorandum of understanding.
2. Statistics Canada will carry out a review of the product and will notify the primary researcher of the acceptance or rejection of the product, including any revisions that may be necessary. Ideally this will happen within two months of the date of submission.
3. Researchers should complete revisions to the product and submit a final draft to Marie Drolet, Project Coordinator at Statistics Canada (613-951-5691 or Marie.Drolet@statcan.ca).
4. Researchers should notify the STC analyst that the project is complete and a final product has been submitted to Statistics Canada. At that time, the researchers must return their security pass/password/identification.
5. Researchers may also choose to save any programming/syntax or output to a CD. This can be done through a request to the STC analyst. Note that these files will be retained for six months following the completion of a Statistics Canada contract and then purged.
6. Researchers are free to publish subsequent reports stemming from their work in the STC.

Note: Your oath of office remains in effect even after you have completed the contract for Statistics Canada.

A8.6 Steps to follow for re-entry of STC user on a new agreement with Statistics Canada

1. Researchers are to submit proposals to STC as they did the first time they wanted access to data. Be sure to include in your proposal a justification for using Statistics Canada microdata. You don't need to re-submit a Curriculum vita if you had done so before.
2. Statistics Canada will carry out a review of the proposal and will notify the primary researcher of the final decision made by the review committee. Ideally this will happen within two months of the date of submission. At that time, Statistics Canada will conduct a security check on all researchers (whom never were subject to security check) and who will be accessing the data in the STC for the first time.
3. Researchers should contact the STC analyst before they would like to access the data and indicate their intent. Upon that contact, four things will happen:
    - The primary researcher will sign a memorandum of understanding between the project team members and Statistics Canada.
    - The researchers will review the orientation material with the STC analyst.
    - The researchers will be asked to reaffirm their oath of office.

    The researchers will then receive their own key/password to access the STC area.

4. Researchers are asked to sign up for a workstation on the days that they would like to access data.
5. Data access begins.

APPENDIX 9

Disclosure Avoidance Guidelines for Using Workplace and Employee Survey Microdata at RDCs

Statistics Canada takes great care to respect the trust of their respondents and to safeguard the privacy and confidentiality of the information that they provide. It is this trust that makes it possible for Statistics Canada to continue to collect accurate and meaningful data. Most household Surveys carried out by Statistics Canada do not require households and business mandatory participation - respondents to volunteer give their time and information freely. The information contained in these and other Statistics Canada surveys benefits the research community, and Statistics Canada goes to great lengths to protect the confidentiality of its respondents' information.

The goal of disclosure avoidance is to protect the information provided by respondents while presenting the least possible hindrance to research. The Statistics Canada staff and researchers will work together to find solutions to confidentiality problems.

Types of data disclosure

Identity disclosure occurs when a specific individual or workplace can be identified from the released data. This type of disclosure is rare but can happen. It ranges from specifically stating whom the respondent is to providing enough information to reveal a respondent's identity. For example, a researcher investigating innovative human resource practices could disaggregate the data to the extent that perhaps only one or two workplaces are contained in a cell (e.g. small unionised workplace in a particular industry with certain human resource practices). Someone who may know most of the characteristics of a given company, particularly if the location of the workplace is revealed, could then easily identify the firm and learn more about it based on the additional information contained in the table.

Attribute disclosure occurs when confidential information is revealed and can be attributed to an individual. For example, if we release the salary range of a particular occupation (e.g. doctors) in a small locality, then there is disclosure if the range gives a better idea of the doctors' salary than would be generally known. Note that in this case we have not identified a particular doctor but, since residents of that locality may know who the people are, identification would occur nonetheless and this amounts to identity disclosure. Note also that we have not given a particular salary figure, but if the range is too narrow, then the salary is assumed to have been revealed. What constitutes 'too narrow a range' may however, be subject to interpretation.

Inferential disclosure occurs when information about an individual can be inferred with a high level of confidence. For example, the results of a regression model may provide a confidence interval for doctors' salaries. In general, statistical agencies do not guard against this type of disclosure because one of the main purposes of statistical data is to enable inferences to be made, and because inferences are not very accurate predictors of individual behaviour.

Residual disclosure occurs when information about a respondent can be detected from the current information and previous information released. This is a particular problem with longitudinal data (e.g., WES) when information is released from subsequent cycles. Alternatively, residual disclosure could occur when information is released from two independent surveys. Residual disclosure may also occur when information in a suppressed cell can be deduced from other information provided. Another type of residual disclosure can occur through sample restrictions for analytical purposes. For example, sample restrictions may exclude some respondents that may be identifiable if compared to all respondents.

Regardless of the process, different types of disclosure are possible but once an individual or firm is identified, identity disclosure has occurred.

All variables on a database can be categorized according to their importance to data confidentiality:

Direct identifiers: Name, address or telephone number provides an explicit link to a respondent. These three variables are stripped from all master files.

Indirect identifiers: Age, sex, marital status, area of residence or occupation, type of business, etc. can be used to identify an individual.

Sensitive variables: These are characteristics relating to respondents' private lives, or business, and are not usually known by the general public.

These variables could work together to reveal information about individuals. Consider the case were indirect identifiers (such as age, sex, marital status and occupation) are presented for a small region along with a sensitive variable such as family income. It may be possible to deduce the family income of certain individuals with a rare combination of these characteristics.

Data confidentiality priorities

Data confidentiality is primarily a problem for frequency data, tables of magnitude and individual statistics. It tends not to be a problem for causal analysis results such as regression parameters.

The following general rules apply at ALL times:

- Outputs have to be checked for confidentiality before they can be taken out of Statistics Canada Offices or the Research Data Centres (RDCs).

- Cross tabulations and charts are discouraged. Cross-tabulations must be vetted for confidentiality prior to leaving the RDC premises and prior to publication. The same applies to charts as they are a graphical representation of cross tabulations.

- No minimum and maximum values can be provided. As well, for highly skewed populations such as earnings, it may be inappropriate to report the 5th and 95^{th} percentile.

- Pay attention to residual disclosure. Residual disclosure may occur when information in a suppressed cell can be deduced from other information provided or when sample restrictions used in the analysis can identify respondents if compared to all respondents.

- Only weighted data can be used for publication. Users are required to provide both unweighted and weighted programs for disclosure analysis. However, only weighted outputs will be released.

- Do not report statistics based on a small number of respondents, which is defined as fewer than 5 cases for the employee data file and fewer than 10 cases for the employer data file. For the employer file, an estimate must be suppressed if 2 or fewer observations contribute to over 90% of that estimate.

- Be aware of certain empty cells and full cells. For example, confidentiality may be broken if the sampled firms in a particular industry and region all reported the same characteristics.

- Anecdotal information should never be given about specific respondents.

- Analytical outputs do not normally present a disclosure problem. However, variables in the model should adhere to the disclosure rules for descriptive statistics and appropriate weights should always be applied.

- Do not report ANOVAs and regression equations when the model involving categorical covariates is saturated or nearly saturated (has many coefficients— intercept, main effects and interaction terms—or nearly as many as there are possible combinations of the covariate values).

The following examples are designed as guidelines for dealing with various data types:

A 9.1 Tabular output: frequency data or tables of magnitude

Data result	Disclosure problem	Solution
Reporting a table of frequencies or magnitudes	Sampling design must be corrected for.	Use weighted data.
Reporting a sample size that represents the sample, not the population	Unweighted sample sizes usually do not pose a confidentiality risk if sample size is greater than 30.	No need to weight data in this case.
Reporting a frequency table or cross-tabulation where a category or cell contains only a few respondents (low frequency cells) Reporting an estimate from a table of magnitude that has a low frequency cell	Reporting small category or cell sizes is a data confidentiality problem and must not be done. Consult the documentation for your survey to determine the definition of a 'small cell size.' Usually it is five.	Collapse categories or exclude categories from analysis.
Reporting a frequency table or cross-tabulation where a category or cell is equal to zero Reporting an estimate from a table of magnitude where a category or cell is equal to zero	There are two kinds of zero cells: 1) structured zero cells, which cannot possibly contain a respondent (e.g., a cell for 'married' and 'under 12 years old'); and 2) non-structural zero cells, which could potentially contain a respondent but do not for a particular analysis.	Structured zero cells are not a data confidentiality problem. Non-structured zero cells should only be published if they account for less than 15% of the non-marginal cells of a table and if they cause no potential disclosure risk; otherwise, collapse categories or exclude categories from the analysis. For a categorical income variable, the zero cells may present a potential disclosure risk if the non-zero cells represent a narrow range of possible values: the highest possible value should not be less than twice the lowest possible value.
Reporting frequency or cross-tabulation tables	The data confidentiality risk depends on the type of	STC staff can provide guidance in deciding when

Data result	Disclosure problem	Solution
where a category or cell contains 100% of the sample (full cell) Reporting an estimate from a table of magnitude that has a full cell	information in the table. There is little risk in publishing full cells when they reveal the sex of respondents. However, it is more problematic when the full cell reveals sensitive information about individuals that would not otherwise be known (i.e., accounting irregularity for all sampled small firms in a particular industry and region).	a full cell proposes a data confidentiality problem. If it has been deemed to be a problem, then collapse categories, exclude categories from analysis, or do an alternative analysis.

Table A9.2 Individual statistics

Data result	Disclosure problem	Solution
Reporting an individual statistic, such as a total, mean, ratio, median or percentile	Sampling design must be corrected for.	Use weighted data.
Reporting a ratio	Ratios should not be published if either component cannot be published.	The ratio should be calculated in another way.
Reporting a total, mean or average based on fewer than three respondents	Reporting statistics from extremely small samples is a data confidentiality problem and must not be done. Consult the documentation for your survey to determine the definition of a 'small sample.' Usually it is three.	Select a bigger sample on which to calculate the statistic.
Reporting order statistics such as medians and percentiles where there are fewer than five respondents above and fewer than five respondents below the order statistic	The 'tails' should contain at least five respondents. If the survey contains multiple respondents from one household, business or organization, then the five respondents should be from at least three different households, businesses, or organizations.	Calculate other order statistics, such as larger percentiles or averages instead of medians.

Table A9.3 Analytical outputs

Data result	Disclosure problem	Solution
Reporting ANOVAs and regression equations	These analytical outputs do not normally present a disclosure problem. Be sure that variables in the model adhere to disclosure rules for descriptive statistics. (See Section A13.5 in this document.)	Should always be calculated on weighted data.
Reporting ANOVAs and regression equations when the model involving categorical covariates is saturated or nearly saturated (has many coefficients—intercept, main effects and interaction terms—or nearly as many as there are possible combinations of the covariate values)	Saturated or nearly saturated models can pose a data confidentiality problem.	Do not calculate saturated or nearly saturated models. Or proceed as when publishing the table whose classification variables are these same covariates, and apply the appropriate rules for tabular outputs. (See Section A13.5 in this document.)
Reporting scatterplots, plots of residuals or box plots	They may present a disclosure risk when they display values for individual respondents, particularly income data with extreme outliers.	Graphical outputs should respect all the rules specified elsewhere in this document.

Table A9.4 Geography and indirect identifiers

Data result	Disclosure problem	Solution
Reporting the location of a sample cluster on a map, list or otherwise	This poses a data confidentiality problem.	Do not do this.
Reporting tabular outputs on variables such as race or ethnicity below the national level	This poses a data confidentiality problem, particularly when there is a great deal of detail for a particularly small geographical area. Exceptions may be granted if the case can be made that revealing more detail is essential to the study report, and does not constitute poor quality data, and does not present a disclosure risk.	Use broad categories such as 'White/Other,' 'English/French/Other,' or 'Canadian/Immigrant.'
Reporting tabular output for, or by, subprovincial areas smaller than 250,000 people Reporting tables that include classification variables that identify very small and/or visible sub-populations Reporting tables that include more than three indirect identifiers as classification variables (in addition to the geographical information)	This can pose a data confidentiality problem.	Apply rules for tabular output
Reporting tables with geographical classification variables (e.g., Health Region, Census Division) or the same geographical classification for two different time periods	This can pose a data confidentiality problem if the table includes more than one geographical classification variable (unless one is an urban/rural code).	Use only one geographical identifier.

Table A9.5 Information about individual respondents

Data result	Disclosure problem	Solution
Reporting maximum or minimum values for sensitive variables such as income, age and household size	This poses a confidentiality problem only when the maximum or minimum value indicates the presence of an atypical respondent.	Report standard deviations or other statistics that can be used to describe the range of values without reporting an actual maximum or minimum.
Reporting anecdotal information about a particular respondent	This is the ultimate confidentiality problem.	Do not do this.

Table A9.6 Related outputs

Data result	Disclosure problem	Solution
Reporting similar information from previous studies or cycles of a survey or from other surveys	This is the most difficult kind of disclosure to control, but every effort should be made to prevent the disclosure of confidential information from related survey data.	Results involving similar sets of classifications (e.g., two types of geographical classification systems, two different 'breakdowns' of occupational codes) should be examined closely. Also, if Public-Use Microdata Files (PUMFs) are released for the same survey, then the published results should not disclose sensitive information that was suppressed from the PUMF about individual respondents.